GMI Computing Update

Tom Clune

SIVO - Code 610.3

Discover

New NCCS Cluster: Discover

- General replacement for alpha cluster (halem)
  - Primary workhorse for GMI for several years
  - Phased installation
    - 1st “base” node - peak of ~ 3.3 TF (halem = 3.2TF)
      - SIVO should get access within 2 weeks
      - General access is scheduled for early November
    - Follow-on nodes >= 2x performance of base node
      - Expect at least 2 such nodes in the next 6 months.
    - Halem to be decommissioned
      - Originally targeted January 1, 2007
      - Power/cooling mandate shutdown prior to 3rd node.
  - IBM GPFS filesystem

Discover Specs

- Linux Networks: Intel/Linux cluster
- 128 nodes connected by Infiniband
  - 5 racks (compare to ~80 racks for halem)
- Dual-core, Dual socket nodes “Demsey”
  - 4 cores/node (1 x halem)
  - 4 GM memory (2 x halem)
  - 3.2 Ghz (~3 x halem)
  - ***Smaller*** cache!
  - *Unproven* scalability within node for NASA apps.

New Data Portal

- Cluster dedicated to serving data in a flexible manner.
  - To be managed by individual projects
  - Little support from NCCS
- Replaces dirac for GMI products?
  - More available than dirac - not dependent on other CXFS platforms
  - Data does not migrate to tape, but must be managed by groups.
  - Can add registration/authentication to process, but probably not in near future.
- Moving beyond FTP?
  - Web access
  - Common visualizations including thumbnails

Processing Met Fields

- Input data
  - 2 forecasts each day
    - Each forecast is 20 snapshots (5 days x 2)
    - Each forecast has ~50 fields (180x288x72)
    - 128 separate files - migrated to varying tapes
  - Aggregate of ~ **250 GB / day**
    - **1 year > 50 TB of data to process!**
- Tape issues
  - At least one phase of processing requires all 128 files from tape.
  - 2-3 hours to assemble 1 days files
  - Some files do not exist.

Processing Met Fields (cont’d)

- Staging data
  - Without intervention data will migrate back to tape before processing completes.
  - Must move data to non-migrating filesystem.
    - Carefully manage available space (~ 1 TB workspace)
- Once data for a single day is staged, actual processing requires ~4 hours for each day.
  - Multiple steps with handcrafted scripts.
  - Differing steps require differing formats - multiple stages to disk
- Overall process is fragile
  - Tape system is unreliable
  - Filesystems fill up
  - Missing data
  - Human error - tedious and time-consuming

Improved workflow

- Partial automation
  - Some steps invoked when data is detected as available
  - Intermediate files are deleted when no longer needed.
  - Automatic registry of completed processing
- Parallelization
  - Multiple files are migrated from tape simultaneously.
  - 5 days processed in batch simultaneously.

Met Fields: Current Status

- Previous performance:
  - ~ 5 days processed per day.
- Current performance:
  - Theoretical ~ 40 days processed per day
  - Actual ~ 20 days processed per day
  - Far less human involvement than before.
- Further improvements?
  - Single coherent script that would require less intermediate representations.
  - Reduced requirement for processing 1 day “all-at-once”.

Miscellaneous Items

- New NCCS allocation request
  - Large number of unused hours on Altix
  - Asked for ~ 50% increase on halem and same number on Altix
    - Will need to request more if new cluster is slow